On estimation of temporal difference error with an exponential filter

Scott Livingston

19 August 2008

The purpose of this draft paper is to address whether one can avoid the "multiple passes" required for learning during each time step when using an artificial neural network for value function approximation (in reinforcement learning) by tracking differences in successive approximations with an exponential filter. This idea is referred to as the "one-pass method" and would allow estimation of the temporal difference error in each time step (assuming discrete time) with only one feedforward operation (which in many cases is still performed without learning, e.g., when a control policy is derived from the action-value function).

Note this is only a draft, hence lacking some appropriate references. For now, an excellent classic text is *Reinforcement Learning: An Introduction* by R. S. Sutton and A. G. Barto (1998; MIT Press). Some detail is left out of the analysis below for the sake of brevity.

Let $\omega \in \mathbb{R}$ such that $0 < \omega < 1$. Call the state set $\mathcal{S}$ and the action set $\mathcal{A}$, and assume both sets are at most countable. Suppose the policy π is deterministic and fixed; accordingly, only policy evaluation is examined here, not policy improvement. Let the reward function r be defined such that it returns a scalar value given a state-action pair (i.e., $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$). Define the action-value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to be the expected return given the policy π and an initial state-action pair. Specifically, given state s_t and action a_t (at time t, as indicated by subscript),

$$Q^\pi(s_t, a_t) = r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \gamma^2 r(s_{t+2}, a_{t+2}) + \cdots .$$

Thus,

$$Q^\pi(s_t, a_t) = r(s_t, a_t) + \gamma Q^\pi(s_{t+1}, a_{t+1}). \tag{1}$$

Denote the approximation of Q^π by $\widetilde{Q}^\pi$. At time t, the difference between Q^π and $\widetilde{Q}^\pi$, referred to as the approximation error at (s_t, a_t), is

$$\begin{aligned} Q^\pi(s_t, a_t) - \widetilde{Q}^\pi(s_t, a_t) &= [r(s_t, a_t) + \gamma Q^\pi(s_{t+1}, a_{t+1})] - [r(s_t, a_t) + \gamma \widetilde{Q}^\pi(s_{t+1}, a_{t+1})] \\ &= \gamma [Q^\pi(s_{t+1}, a_{t+1}) - \widetilde{Q}^\pi(s_{t+1}, a_{t+1})] \end{aligned}$$

In the one-pass method, we have

$$\rho_t = (1 - \omega)\rho_{t-1} + \omega[\widetilde{Q}^\pi(s_t, a_t) - \widetilde{Q}^\pi(s_{t-1}, a_{t-1})] \tag{2}$$

and the next action-value approximation (of an approximation),

$$\widehat{\widetilde{Q}}^{\pi}(s_{t+1}, a_{t+1}) = \widetilde{Q}^{\pi}(s_t, a_t) + \rho_t.$$

The approximation error for the one-pass method is thus

$$\begin{aligned} Q^{\pi}(s_t, a_t) - \widetilde{Q}^{\pi}(s_t, a_t) &= \gamma[Q^{\pi}(s_{t+1}, a_{t+1}) - \widehat{\widetilde{Q}}^{\pi}(s_{t+1}, a_{t+1})] \\ &= \gamma[Q^{\pi}(s_{t+1}, a_{t+1}) - \widetilde{Q}^{\pi}(s_t, a_t) - \rho_t] \qquad (3) \end{aligned}$$

To better understand what convergence may look like, assuming it is possible, I consider the case when approximation error is zero for all state-action pairs and investigate its implications. The left-hand side of equation (3) is thus 0, and we have

$$Q^{\pi}(s_{t+1}, a_{t+1}) = Q^{\pi}(s_t, a_t) + \rho_t,$$

and perhaps more interestingly,

$$\rho_t = Q^{\pi}(s_{t+1}, a_{t+1}) - Q^{\pi}(s_t, a_t). \qquad (4)$$

Now I will show that this implies that the running average ρ_t must be constant. Beginning with equation (2),

$$\begin{aligned} \rho_t &= (1-\omega)\rho_{t-1} + \omega[\widetilde{Q}^{\pi}(s_t, a_t) - \widetilde{Q}^{\pi}(s_{t-1}, a_{t-1})] \\ &= (1-\omega)\rho_{t-1} + \omega[Q^{\pi}(s_t, a_t) - Q^{\pi}(s_{t-1}, a_{t-1})] \qquad \text{because } Q^{\pi}(s,a) - \widetilde{Q}^{\pi}(s,a) = 0 \quad \forall s, a \\ &= (1-\omega)\rho_{t-1} + \omega\rho_{t-1} \qquad \text{by equation (4)} \\ &= \rho_{t-1} \end{aligned}$$

Therefore, $\rho_t = \rho_0$ for all time steps $t \geq 0$. Because it has been shown that the running average ρ_t is constant, it will now be written as ρ.

Equation (4) now indicates that, given the hypothesis of action-value function convergence, the difference between each consecutive state-action pair under the policy π is constant.

One final interesting observation, again under the hypothesis of convergence, is made by combining equations (1) and (4),

$$\begin{aligned} Q^{\pi}(s_t, a_t) &= Q^{\pi}(s_{t+1}, a_{t+1}) - \rho \\ &= \frac{1}{\gamma}\left[Q^{\pi}(s_t, a_t) - r(s_t, a_t)\right] - \rho, \end{aligned}$$

and then rearranging terms to arrive at

$$Q^{\pi}(s_t, a_t) = \frac{\gamma\rho + r(s_t, a_t)}{1-\gamma}. \qquad (5)$$

This finding is important because it follows from the hypothesis of convergence; that is, equation (5) is *necessary* for convergence.

It is still not immediately clear what structure of the environment (a Markov decision process) is necessary for convergence. By combining Eqs. (4) (recall that ρ is constant) and (5), we have

$$\begin{aligned}\rho &= \frac{\gamma\rho + r(s_{t+1}, a_{t+1})}{1-\gamma} - \frac{\gamma\rho + r(s_t, a_t)}{1-\gamma} \\ &= \frac{r(s_{t+1}, a_{t+1}) - r(s_t, a_t)}{1-\gamma}.\end{aligned}$$

Rearranging terms leads to

$$r(s_{t+1}, a_{t+1}) = r(s_t, a_t) + \rho(1-\gamma), \tag{6}$$

which is a recursive definition for the reward function along any trajectory under policy π. A trajectory is a sequence of state-action pairs, $\langle (s_0, a_0), (s_1, a_1), \ldots \rangle$. Eq. (6) can be solved to find the reward at time t (recall that time is discretized such that $t \in \{0, 1, 2, \ldots\}$):

$$r(s_t, a_t) = t\rho(1-\gamma) + r(s_0, a_0). \tag{7}$$